Distribution-Preserving Statistical Disclosure Limitation[1]

Simon D. Woodcock[2]
Simon Fraser University
simon_woodcock@sfu.ca

Gary Benedetto
US Census Bureau
gary.linus.benedetto@census.gov

May 24, 2009

[1]This document reports the results of research and analysis undertaken by the U.S. Census Bureau staff. This document is released to inform interested parties of research and to encourage discussion. This research is a part of the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD), which is partially supported by the National Science Foundation Grants SES-9978093 and SES-0427889 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging Grant R01 AG018854, and the Alfred P. Sloan Foundation. The views expressed on statistical, methodological, or technical issues are those of the authors and not necessarily those of the U.S. Census Bureau, its program sponsors or data providers. Some or all of the data used in this paper are confidential data from the LEHD Program. The U.S. Census Bureau supports external researchers' use of these data through the Research Data Centers (see www.ces.census.gov). For other questions regarding the data, please contact Jeremy S. Wu, Program Manager, U.S. Census Bureau, LEHD Program, Demographic Surveys Division, FOB 3, Room 2138, 4700 Silver Hill Rd., Suitland, MD 20233, USA. (Jeremy.S.Wu@census.gov http://lehd.dsd.census.gov).

[2]Correspondence to: Simon Woodcock, Department of Economics, Simon Fraser University, 8888 University Dr., Burnaby, BC V5A 1S6, Canada. We thank John Abowd, Sam Hawala, Pascal Lavergne, Krishna Pendakur, members of the LEHD Program staff, and members of the Canadian Econometric Study Group for helpful comments and suggestions. We also thank Bryan Richetti for providing SAS code used in the re-identification analysis. We gratefully acknowledge the financial support of NSF grants ITR-0427889, EITM-0339191, SES-9978093 awarded to Cornell University, NIA grant 5R01AG018854-03, the SSHRC institutional grants program, the SFU President's Research Grant, and the SFU Community Trust Endowment Fund.

Abstract

One approach to limiting disclosure risk in public-use microdata is to release multiply-imputed, partially synthetic data sets. These are data on actual respondents, but with confidential data replaced by multiply-imputed synthetic values. A mis-specified imputation model can invalidate inferences based on the partially synthetic data, because the imputation model determines the distribution of synthetic values. We present a practical method to generate synthetic values when the imputer has only limited information about the true data generating process. We combine a simple imputation model (such as regression) with density-based transformations that preserve the distribution of the confidential data, up to sampling error, on specified subdomains. We demonstrate through simulations and a large scale application that our approach preserves important statistical properties of the confidential data, including higher moments, with low disclosure risk.

Keywords: statistical disclosure limitation, confidentiality, privacy, multiple imputation, partially synthetic data

1 Introduction

Statistical agencies face two competing objectives when preparing data for public release. On the one hand, they endeavor to provide their users with high quality data. On the other hand, they must maintain the privacy of respondents. There is a trade-off between these objectives because protecting privacy usually entails information loss (Duncan et al., 2001). Unless care is taken, measures to protect privacy can invalidate statistical inferences.

One way to protect privacy in public-use microdata is to release multiply-imputed, partially synthetic data sets. These are data on actual respondents, but with confidential data replaced by multiply-imputed synthetic values. When the imputation model is correctly specified, the multiply-imputed partially synthetic data permit valid inferences about the population of interest. However, a mis-specified imputation model can invalidate inferences, because the distribution of synthetic data is determined by the model that generates them.

We present a practical method to generate synthetic values when the imputer has only limited information about the true data generating process. We combine a simple imputation model (such as regression) with density-based transformations that preserve the distribution of the confidential data, up to sampling error, on specified subdomains. This allows users to obtain valid inferences about a variety of quantities, with low disclosure risk. We demonstrate this via simulations and an application to the US Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) database.

Traditional approaches to limit disclosure risk include suppressing confidential data, aggregation, topcoding, adding noise, and swapping values between records (see e.g., Willenborg and de Waal (1996) or the appendix to Abowd and Woodcock (2001) for surveys). All of these have the potential to distort the joint distribution of the data, and may therefore invalidate inference. In many cases, valid inferences can only be obtained using specialized software and methods, and/or when users are provided with detailed information about the methods used to limit disclosure risk. In practice, however, such detailed information cannot be released without compromising privacy.

An alternative that permits valid statistical inferences using standard software and methods is to release data sets comprised entirely of synthetic values sampled from an estimate of the joint distribution of the confidential database. Rubin (1993) proposes multiple imputation to generate the synthetic values;[1] Fienberg (1994) suggests bootstrap methods.[2] Under either approach, the released data pose little or no disclosure risk because they are completely synthetic, i.e., contain no actual data on actual respondents. However, this approach requires knowledge, or a good estimate, of the joint distribution of the data. This is impractical in many instances. A tractable alternative is to release data on actual respondents, but replace confidential data with synthetic values sampled from an estimate of the joint distribution of the confidential data conditional on disclosable data. Such data, which have become known as partially synthetic data, are the focus of this paper.

Kennickell (1997) pioneered the use of multiply-imputed, partially synthetic data in the Survey of Consumer Finances. Subsequent authors have suggested several approaches to generate the synthetic values. Abowd and Woodcock (2001) propose a computationally tractable approximation to the joint distribution of the confidential data given disclosable data based on a sequence of regression models. They use this approximation to multiply-impute confidential values in linked employer-employee data. Little and Liu (2003) develop a parametric method, called SMIKe, to selectively multiply-impute discrete "key" variables that pose high disclosure risk. Reiter (2005d) proposes a nonparametric method to multiply-impute synthetic values using classification and regression trees (CART).

Each of these approaches makes an important contribution, but all have limitations. SMIKe is only applicable to categorical key variables. CART, though data-driven and requiring little modeling input from the imputer, may be more difficult to describe to the public than a parametric model. Simple descriptions of the imputation model are useful meta-data for public-use releases, since they help users determine which analyses can be reasonably

[1] This proposal is developed more fully in Raghunathan et al. (2003). Reiter (2002) provides a simulation study, Reiter (2005c) discusses inference, and Reiter (2005b) provides an application.

[2] Fienberg et al. (1998) apply this method to categorical data; Fienberg and Makov (1998) use related concepts to develop a measure of disclosure risk.

supported by the synthetic data. And though Abowd and Woodcock (2001) demonstrate that regression-based methods perform well in practice, the regression models are subject to mis-specification when the true data generating process is unknown.

Our approach is predicated on the assumption that data collectors prefer to use simple (or convenient) imputation models to generate synthetic values, such as regression models. We believe this assumption reflects reality at many statistical agencies. Data collectors may prefer simple imputation models for various reasons: to reduce modeling or computational burden, because they are easy to diagnose and interpret, or because the correct imputation model is unknown. However, synthetic data generated using a simple imputation model may fail to reproduce complex features of the confidential data, such as nonlinear relationships between variables, skewness, tail thickness, and the number and location of modes. All of these may be important for obtaining valid inferences about some quantities.

Our proposed approach is similar in spirit to the nonlinear data-fitting methods of Lin and Vonesh (1989) and Nusser et al. (1996), and the copula-based additive noise perturbation of Sarathy et al. (2002). It proceeds as follows. First, we divide the data into subdomains of primary interest. Second, in each subdomain, we transform the variable under imputation to have a standard distribution that is compatible with a simple imputation model. Then we generate synthetic values using a simple model on the transformed data. The role of the simple imputation model is to preserve relationships of secondary interest within subdomains. Finally, we apply an inverse transformation that returns the synthetic values to the native scale and distribution of the underlying confidential variable. This preserves the distribution of the confidential variable on the subdomains of primary interest.

Our approach is less subject to mis-specification than a simple imputation model alone, because we only rely on the simple model to capture relationships of secondary interest. In fact, our simulations and application demonstrate that our approach preserves important statistical properties of the confidential data, including higher moments, with low disclosure risk. Furthermore, it is easily applied in practical situations involving many variables and

observations.

The remainder of the paper is organized as follows. We begin, in Section 2, by reviewing some relevant background information. We develop our synthesis method in Section 3. Section 4 presents the simulations, Section 5 presents the application to linked employer-employee data, and Section 6 concludes.

2 Background and Concepts

Consider a database that consists of confidential microdata $\mathbf{Y}$ and disclosable microdata $\mathbf{X}$. Both $\mathbf{X}$ and $\mathbf{Y}$ may contain discrete and continuous elements. Let $\mathbf{D} = (\mathbf{X}, \mathbf{Y})$ represent the database in question, and $F(\mathbf{D})$ its probability distribution.

The data collector wishes to release public microdata $\tilde{\mathbf{D}}$. Her competing objectives are to maximize data utility and minimize disclosure risk. Muralidhar and Sarathy (2003) suggest a very stringent criterion for data utility: that the observed and released data are identically distributed, $F(\tilde{\mathbf{D}}) = F(\mathbf{D})$. In practice, this cannot be achieved because F is usually unknown and because disclosure limitation entails information loss. Instead, usual practice is to require that the released data yield valid inferences about quantities of substantive interest, e.g., means, variances, regression coefficients, and the like. Disclosure risk is usually assessed via simulations that attempt to replicate behavior of a malicious data user (i.e., an intruder or snooper) who seeks to infer the value of a confidential datum.[3]

2.1 Multiply-Imputed, Partially Synthetic Data

Partially synthetic data replace confidential values $\mathbf{Y}$ with synthetic values $\tilde{\mathbf{Y}}$**.** A partially synthetic data release is $\tilde{\mathbf{D}} = (\mathbf{X}, \tilde{\mathbf{Y}})$. Multiply-imputed, partially synthetic (MIPS) data rely on an imputation model to generate the synthetic values. In the parametric case, this

[3]Duncan and Lambert (1986) and Lambert (1993) propose a general framework to assess the risk of identity disclosure. They model the behavior of an intruder to obtain disclosure probabilities and Bayesian measures of uncertainty about those probabilities. Reiter (2005a) demonstrates their approach for several traditional disclosure limitation methods.

is defined by a likelihood $p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta})$ and prior $p(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are unknown parameters. Synthetic values are sampled from the posterior predictive distribution:

$$p\left(\tilde{\mathbf{Y}}|\mathbf{X},\mathbf{Y}\right) = \int p\left(\tilde{\mathbf{Y}}|\mathbf{X},\mathbf{Y},\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}|\mathbf{X},\mathbf{Y}\right) d\boldsymbol{\theta}. \tag{1}$$

The likelihood in (1) conditions on $\mathbf{Y}$ to reflect the possibility that we selectively impute values (e.g., when they exceed a threshold), or use different imputation models on subdomains of $\mathbf{Y}$. To simplify notation in what follows, we usually omit $\mathbf{Y}$ from the conditioning statement in the likelihood.

Specifying the joint likelihood can be challenging in practice. This is particularly true when there are many confidential variables, when some are continuous and others are discrete, and when relationships among variables are complex. This is often the case in genuine applications. Specifying the joint likelihood as a sequence of univariate conditional likelihoods reduces modeling burden somewhat. If we write $\mathbf{Y} = [\mathbf{y}_1 \;\; \mathbf{y}_2 \;\; \cdots \;\; \mathbf{y}_K]$ and $\boldsymbol{\theta} = [\boldsymbol{\theta}_1 \;\; \boldsymbol{\theta}_2 \;\; \cdots \;\; \boldsymbol{\theta}_K]$, we can use the factorization

$$p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta}) = p_1(\mathbf{y}_1|\mathbf{X},\boldsymbol{\theta}_1)\, p_2(\mathbf{y}_2|\mathbf{X},\mathbf{y}_1,\boldsymbol{\theta}_2) \cdots p_K(\mathbf{y}_K|\mathbf{X},\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_{K-1},\boldsymbol{\theta}_K) \tag{2}$$

based on a univariate likelihood for each $\mathbf{y}_k$.[4] This factorization accommodates continuous and discrete variables by choice of likelihood for each $\mathbf{y}_k$, and admits complex relationships between variables via conditional dependence. Factorization also lets us generate synthetic values $\tilde{\mathbf{Y}} = [\tilde{\mathbf{y}}_1 \;\; \tilde{\mathbf{y}}_2 \;\; \cdots \;\; \tilde{\mathbf{y}}_K]$ sequentially, e.g., by sampling $\tilde{\mathbf{y}}_1$ from the posterior predictive distribution of $\mathbf{y}_1$ given $\mathbf{X}$, then sampling $\tilde{\mathbf{y}}_2$ from the posterior predictive distribution of $\mathbf{y}_2$ given $\mathbf{X}$ and $\tilde{\mathbf{y}}_1$, and so on.

It is well known that multiple imputation yields valid inferences when the imputation

[4] An alternative, proposed by Abowd and Woodcock (2001) and based on the Sequential Regression Multivariate Imputation (SRMI) algorithm of Raghunathan et al. (2001), is to approximate the joint likelihood by a sequence of regression models. This is an iterative procedure, consisting of L rounds of synthesis. In each round, synthetic values are drawn sequentially for each $\mathbf{y}_k$, conditional on $\mathbf{X}$ and the most recently-drawn synthetic values for all other confidential variables.

model is correctly specified. We call one draw from the posterior predictive distribution a partially synthetic data implicate, $\tilde{\mathbf{Y}}^m$. The data collector must release multiple implicates, $\tilde{\mathbf{D}}^m = \left(\mathbf{X}, \tilde{\mathbf{Y}}^m\right)$ for $m = 1, 2, ..., M$, to allow valid inferences from the MIPS data. Reiter (2003) develops the necessary distribution theory.[5] Suppose that with access to the confidential data $\mathbf{D}$, users would base inference about a scalar population quantity Q on a sample statistic q with asymptotic distribution $(Q - q) \overset{a}{\sim} N(0, V)$. The user computes the sample statistic q^m on each partially synthetic data implicate. Let v^m denote the sampling variance of q^m. Estimates from the M implicates are combined using:

$$\bar{q}_M = \frac{1}{M}\sum_{m=1}^{M} q^m, \quad b_M = \frac{1}{M-1}\sum_{m=1}^{M}\left(q^m - \bar{q}_M\right)^2, \quad \bar{v}_M = \frac{1}{M}\sum_{m=1}^{M} v^m. \tag{3}$$

An unbiased estimator of the variance of $\bar{q}_M$ is $T = M^{-1}b_M + \bar{v}_M$. In large samples, inferences about Q can be based on a t distribution with $\nu = (M-1)\left(1 + r_M^{-1}\right)^2$ degrees of freedom, where $r_M = (M^{-1}b_M/\bar{v}_M)$. These combining rules differ slightly from those for multiply-imputed missing data (e.g., Rubin, 1987). To understand this, note that in the absence of missing data, $\bar{v}_M$ estimates the variance of $Q|\mathbf{D}$. In contrast, in standard multiple imputation for missing data, $\bar{v}_M + b_M$ estimates the variance of $Q|\mathbf{D}$. For an extended discussion of this point, see Reiter and Raghunathan (2007).

3 Imputation Using Simple Models and Transformations

A mis-specified imputation model can invalidate inference based on MIPS data.[6] This is significant in genuine applications, because data collectors are unlikely to know $F(\mathbf{D})$. The synthesis procedure we develop here is designed to mitigate mis-specification that arises when

[5] Reiter (2004) considers the case where multiple imputation is used both for missing data imputation and disclosure limitation.

[6] A mis-specified imputation model could also affect disclosure risk, but there is no particular reason to expect it will increase.

the correct imputation model is unknown.

Suppose the data collector wishes to generate synthetic values of a continuous variable $\mathbf{y}_k$ conditional on a subset of information in the database, $\mathbf{W} \subseteq \mathbf{D}$. Define a partition of the conditioning information $\mathbf{W} = [\mathbf{W}_1 \ \ \mathbf{W}_2]$ such that relationships between $\mathbf{y}_k$ and $\mathbf{W}_1$ are of primary interest, and relationships between $\mathbf{y}_k$ and $\mathbf{W}_2$ are of secondary interest. The objective is to replicate the distribution of $\mathbf{y}_k$ on subdomains of $\mathbf{W}_1$, and preserve key relationships between $\mathbf{y}_k$ and $\mathbf{W}_2$ on those subdomains. We have in mind that $\mathbf{W}_1$ describes a collection of discrete variables, or a meaningful discretization of continuous variables (e.g., categories based on quantiles). For example, suppose that $\mathbf{y}_k$ is employment income and $\mathbf{W}_1$ is a collection of discrete characteristics such as sex, race, geography, and educational attainment. The objective is to preserve the distribution of income in sex $\times$ race $\times$ geography $\times$ education cells, and to preserve key relationships between income and $\mathbf{W}_2$ (e.g., age, employer size, etc.) within those cells.

We assume the data collector does not know the correct model for $\mathbf{y}_k|\mathbf{W}_2$ on subdomains of $\mathbf{W}_1$. She therefore favors a simple (or convenient) imputation model to generate synthetic values, such as regression, that is easy to diagnose and interpret, and that captures important relationships between $\mathbf{y}_k$ and $\mathbf{W}_2$. This poses two problems. First, the distribution of synthetic values sampled from the posterior predictive distribution of a simple imputation model for $\mathbf{y}_k|\mathbf{W}_2$ may not coincide with the observed distribution of $\mathbf{y}_k$ on the subdomain, $F_{y|W_1=w_1}$. This distorts the distribution of $\mathbf{y}_k|\mathbf{W}_1$ in the synthetic data and may invalidate inference. Second, incompatibility between the distribution of $\mathbf{y}_k$ and the simple imputation model may exacerbate mis-specification. For example, a linear regression of $\mathbf{y}_k$ on $\mathbf{W}_2$ may fit poorly and yield implausible synthetic values if $F_{y|W_1=w_1}$ is highly skewed or multi-modal, distorting relationships between $\mathbf{y}_k$ and $\mathbf{W}_2$. Our solution is to transform $\mathbf{y}_k$ in a way that improves the fit of the simple imputation model, and apply an inverse transformation to the synthetic values that replicates the distribution of $\mathbf{y}_k$ on the subdomain.

Let $p\left(\mathbf{z}_k|\mathbf{W}_2, \mathbf{W_1} = \mathbf{w_1}, \boldsymbol{\theta}_k\right)$ denote the likelihood of a simple imputation model for

$\mathbf{z}_k|\mathbf{W}_2$ on a subdomain of $\mathbf{W}_1$, where $\mathbf{z}_k$ is a monotone transformation of $\mathbf{y}_k$ that improves the model's fit. The transformation could be deterministic (such as a logarithm) or stochastic. We focus on a general-purpose density-based transformation.

Let $F_{z|W_1=w_1}$ denote a distribution function compatible with the simple imputation model, e.g., the distribution function implied by the likelihood averaged over $\mathbf{W}_2$ and $\boldsymbol{\theta}_k$. Define the transformation

$$\mathbf{z}_k \equiv F^{-1}_{z|W_1=w_1}\left(\hat{F}_{y|W_1=w_1}\left(\mathbf{y}_k|\mathbf{W}_1=\mathbf{w}_1\right)\right) \tag{4}$$

where $\hat{F}_{y|W_1=w_1}$ is an estimate of $F_{y|W_1=w_1}$. Now $\mathbf{z}_k \sim F_{z|W_1=w_1}$ by construction. Because $\mathbf{z}_k$ is a monotone transformation of $\mathbf{y}_k$, this transformation preserves monotone and rank-order relationships between $\mathbf{y}_k$ and $\mathbf{W}_2$.

Let $\tilde{\mathbf{z}}_k$ denote synthetic values sampled from the posterior predictive distribution:

$$p\left(\tilde{\mathbf{z}}_k|\mathbf{W}_2,\mathbf{W_1}=\mathbf{w_1},\mathbf{z}_k\right)=\int p\left(\tilde{\mathbf{z}}_k|\mathbf{W}_2,\mathbf{W_1}=\mathbf{w_1},\boldsymbol{\theta}_k\right)p\left(\boldsymbol{\theta}_k|\mathbf{W}_2,\mathbf{W_1}=\mathbf{w_1},\mathbf{z}_k\right)d\boldsymbol{\theta}_k. \tag{5}$$

The synthetic values are distributed $\tilde{\mathbf{z}}_k \sim F_{\tilde{z}|W_1=w_1}$ on the subdomain, where $F_{\tilde{z}|W_1=w_1}$ is defined by the predictive distribution. In some cases $F_{\tilde{z}|W_1=w_1}$ will be known. More generally, we will need to estimate it. Let $\hat{F}_{\tilde{z}|W_1=w_1}$ denote an estimate. Define synthetic values $\tilde{\mathbf{y}}_k$ by the inverse transformation:

$$\tilde{\mathbf{y}}_k = \hat{F}^{-1}_{y|W_1=w_1}\left(\hat{F}_{\tilde{z}|W_1=w_1}\left(\tilde{\mathbf{z}}_k\right)\right). \tag{6}$$

Because $\tilde{\mathbf{y}}_k$ is a monotone transformation of $\tilde{\mathbf{z}}_k$, this transformation preserves monotone and rank-order relationships between synthetic values and $\mathbf{W}_2$. Moreover, the synthetic values are distributed according to $\tilde{\mathbf{y}}_k \sim \hat{F}_{y|W_1=w_1}$ by construction. Repeating this procedure on each subdomain of $\mathbf{W}_1$ yields synthetic data that replicate the distribution of $\mathbf{y}_k|\mathbf{W}_1$, up to sampling error in $\hat{F}_{y|W_1=w_1}$.

To obtain valid inferences from the synthetic data, the imputations must be *proper* in the

sense of Rubin (1987), i.e., propagate model uncertainty across the implicates. Our transformations (4) and (6) contain sampling error because they are based on sample estimates. It is important to introduce between-implicate variation in the estimated transformations to reflect this uncertainty. One solution is to estimate $\hat{F}_{y|W_1=w_1}$ and $\hat{F}_{\tilde{z}|W_1=w_1}$ on an approximate Bayesian bootstrap sample of observations in each implicate. The estimates could be based on the empirical distribution function or a smoothed estimate. The latter will provide better protection against attribute disclosure, because observed values are not replicated exactly in the synthetic data.

The partition of $\mathbf{W}$ into $\mathbf{W}_1$ and $\mathbf{W}_2$ is arbitrary, but there is a trade-off. Including more elements in $\mathbf{W}_1$ preserves more dimensions of the distribution of $\mathbf{y}|\mathbf{W}$. However, it also reduces the number of observations in each subdomain, thereby reducing the precision of the estimated transformations and of the simple imputation model. It also increases computational burden, since the density and simple imputation model are estimated on each subdomain.[7]

Preserving relationships of secondary interest between $\mathbf{y}_k$ and $\mathbf{W}_2$ requires a well-specified imputation model for $\mathbf{z}_k|\mathbf{W}_2$ in each subdomain. Our procedure does not alleviate the modeling burden for $\mathbf{z}_k|\mathbf{W}_2$. We have in mind simple but flexible imputation models, e.g., linear regression on polynomials of predictors in $\mathbf{W}_2$ and interactions between them, that capture relationships of interest between $\mathbf{y}_k$ and $\mathbf{W}_2$. As demonstrated in the simulations and empirical application below, this is sufficient to obtain valid inferences about a variety of quantities, including means, variances, correlations, and regression coefficients.

Our approach has some advantages over semiparametric and nonparametric alternatives for generating synthetic values on each subdomain. Sampling synthetic values from a nonparametric estimate of the conditional distribution of $\mathbf{y}_k|\mathbf{W}_2$, or from the predictive distribution of a nonparametric regression, suffers the usual problems of nonparametric estimation:

[7]Either $\mathbf{W}_1$ or $\mathbf{W}_2$ could be empty. When $\mathbf{W}_2$ is empty, synthetic values $\tilde{\mathbf{y}}_k$ can be resampled from the observed data on each subdomain of $\mathbf{W}_1$. If this does not provide sufficient disclosure protection, synthetic values can be sampled from a smoothed estimate of $F_{y|W_1=w_1}$.

they suffer the curse of dimensionality and are computationally costly. Alternatives with more modest computational requirements, such as spline regression, impose restrictions on modeled relationships (such as additive separability) that may distort relationships between $\mathbf{y}_k$ and $\mathbf{W}_2$. Relaxing these restrictions, e.g., including interaction terms between predictors, may rapidly increase computational burden. Furthermore, there is no guarantee that the distribution of synthetic values will replicate $F_{y|W_1=w_1}$, invalidating some inferences. In contrast, our approach preserves distributional features of primary interest, is easy to implement in standard software, computationally cheap, and numerically stable. Furthermore, a simple parametric imputation model for $\mathbf{z}_k|\mathbf{W}_2$ will be sufficient when the data collector knows which relationships she wants to preserve in the synthetic data, and when these are well described by a parametric model.

Example 1 *The data collector would like to use a linear regression model to generate synthetic values of* $\mathbf{y}_k$ *conditional on* $\mathbf{W}$*, but the distribution of* $\mathbf{y}_k|\mathbf{W}$ *is not normal. Let* $\mathbf{W}_1$ *be a collection of categorical variables in* $\mathbf{W}$*, and* $\mathbf{W}_2$ *the remaining elements of* $\mathbf{W}$*. Define the subdomains* $\mathbf{w}_1$ *according to the cells of the cross-classification of variables in* $\mathbf{W}_1$*. On each subdomain, estimate* $\hat{F}_{y|W_1=w_1}$ *on an approximate Bayesian bootstrap sample of observations. Define the transformed values* $\mathbf{z}_k = \Phi^{-1}\left(\hat{F}_{y|W_1=w_1}\left(\mathbf{y}_k|\mathbf{W}_1=\mathbf{w}_1\right)\right)$*, where* Φ *denotes the standard normal CDF, so that* $\mathbf{z}_k \sim N(0,1)$ *on each subdomain. Up to location and scale, this is the marginal distribution implied by the likelihood of the imputation model when* $\mathbf{W}_2$ *are multivariate normal. More generally, the transformation to* $\mathbf{z}_k$ *improves fit of the regression model by rendering the distribution symmetric and unimodal. Next, sample synthetic values* $\tilde{\mathbf{z}}_k$ *from the posterior predictive distribution defined by the normal linear regression of* $\mathbf{z}_k$ *on* $\mathbf{W}_2$ *and an uninformative prior. The synthetic values* $\tilde{\mathbf{z}}_k$ *are distributed* t *with center* $\mathbf{W}_2\hat{\boldsymbol{\beta}}$ *and* $n-k$ *degrees of freedom. Averaged over* $\mathbf{W}_2$ *and parameters, their distribution is approximately standard normal; so define the inverse transformation* $\tilde{\mathbf{y}}_k = \hat{F}^{-1}_{y|W_1=w_1}\left(\Phi\left(\tilde{\mathbf{z}}_k\right)\right)$*. If the standard normal approximation is poor, construct a sample estimate of* $F_{\tilde{z}|W_1=w_1}$ *from the* $\tilde{\mathbf{z}}_k$*. Denote the estimate* $\hat{F}_{\tilde{z}|W_1=w_1}$ *and define the synthetic values* $\tilde{\mathbf{y}}_k = \hat{F}^{-1}_{y|W_1=w_1}\left(\hat{F}_{\tilde{z}|W_1=w_1}\left(\tilde{\mathbf{z}}_k\right)\right)$*.*

In either case, the synthetic and confidential values are identically distributed (up to sampling error) on the subdomain $\mathbf{W}_1 = \mathbf{w}_1$ *i.e.,* $\tilde{\mathbf{y}}_k \sim \hat{F}_{y|W_1=w_1}$. *Repeating this procedure* M *times on each subdomain yields* M *synthetic imputations that replicate the distribution of* $\mathbf{y}_k|\mathbf{W}_1$.

4 Simulation

We illustrate and evaluate our synthesis method with a brief simulation. The objectives of the simulation are threefold. First, to assess performance of our method relative to the case where the exact imputation model is known. Second, to compare its performance to a nonparametric alternative: additive spline regression on each subdomain. And third, to assess identity disclosure risk in the partially synthetic data, since we are unable to do so in our empirical application (Section 5).

We simulate 5,000 databases, each comprising 10,000 observations on six variables. Of the six variables, we treat three as disclosable and three as confidential. We generate three partially synthetic implicates of each simulated database.

The disclosable variables are defined as follows. The first, denoted g, takes value one or two with equal probability. We refer to g as an observation's *group*. The other disclosable variables are x_1 and x_2, independently distributed $N(0,1)$ and rounded to the nearest integer on $[-2,2]$. The confidential variables are defined as follows:

$$y_1 = \exp\left\{3g + \left(g^{1/2}/3\right)x_1 + \left(g^{1/2}/3\right)x_2 + \varepsilon_1\right\} \tag{7}$$

$$y_2 = \exp\left\{3g + \left(g^{1/2}/4\right)x_1 + \left(g^{1/2}/4\right)x_2 + \left(g^{1/2}/4\right)\ln(y_1) + \varepsilon_2\right\} \tag{8}$$

$$y_3 = F^{-1}_{y_3|g}\left(F_{z_3|g}(z_3)\right) \tag{9}$$

where $z_3 = x_1 - (g/2)^{1/2} x_2 + \varepsilon_3$; the errors are independently distributed $\varepsilon_1 \sim N(0, g/9)$, $\varepsilon_2 \sim N(0, g/16)$, and $\varepsilon_3 \sim N(0, g/2)$; and where $F_{y_3|g}$ is the cdf of a $70:30$ mixture of a $N(g, g^2)$ and a $N(3g, g^2/4)$, and $F_{z_3|g}$ is the cdf of $z_3|g$.[8] Conditional on g, the distributions of y_1 and

[8] Note $z_3|g \sim N(0, 1+g)$, $z_1|g \sim N(3g, g/3)$, and $z_2|g \sim N\left(3g\left[1+g^{1/2}/4\right], g\left[3+g/3+4g^{1/2}/3\right]/16\right)$.

y_2 are highly skewed and that of y_3 is bimodal. Subject to the monotone transformations $z_1 = \ln(y_1)$, $z_2 = \ln(y_2)$, and $z_3 = F^{-1}_{z_3|g}\left(F_{y_3|g}(y_3)\right)$, however, they are conditionally and marginally normal in each group.

Conditional on disclosable variables, y_3 is independent of y_1 and y_2. Hence we synthesize y_3 independently of the other confidential variables, with $\mathbf{W}_1 = g$ and $\mathbf{W}_2 = \{x_1, x_2\}$. We synthesize y_1 and y_2 sequentially: y_1 first, with $\mathbf{W}_1 = g$ and $\mathbf{W}_2 = \{x_1, x_2\}$, and y_2 second with $\mathbf{W}_1 = g$ and $\mathbf{W}_2 = \{x_1, x_2, y_1\}$. We generate synthetic data for all three confidential variables following Example 1, using integrated kernel densities to define the transformations. We also apply the density-based transformation to the predictor y_1 in the model for y_2.

To assess the information loss due to ignorance of the exact transformation between the confidential variables and the correct imputation model, we also synthesize y_1 and y_2 using the exact (correct) imputation model: linear regression in logarithms on each subdomain. We also generate synthetic data for all three variables using additive spline regressions. On each subdomain of $\mathbf{W}_1$, we specify an imputation model with additively-separable cubic smoothing splines for each variable in $\mathbf{W}_2$. More general spline regression specifications (e.g., with interactions) encountered computational difficulties.[9] To ensure the synthetic data reflect model uncertainty, we fit the spline regressions to an approximate Bayesian bootstrap sample in each implicate. We then calculate a predicted value and residual for each record and generate synthetic values by adding a bootstrapped residual, resampled from a donor observation in the same decile, to the predicted value.[10]

Figure 1 summarizes bias in the first four moments and selected quantiles of the distribution of synthetic data. For each of the three synthetic variables, we calculate relative bias in each moment and quantile by group, and plot the average bias in 5000 replications.

[9] Not reported, but available on request, are simulation results for synthesis based on local regression, also known as scatter-plot smoothing or LOESS. Local regression proved computationally more robust than additive splines in specifications with interactions, but increased computational time twenty-five fold with no appreciable improvement in synthetic data quality versus reported results for spline regression.

[10] Reported results for spline regression exclude a small number of replications that encountered computational difficulties. The spline estimator apparently failed to converge in these replications. We believe this problem is specific to the SAS GAM procedure, which is deemed experimental in SAS version 9.

For moments, relative bias is $(\bar{q}_3 - Q)/Q$ where $\bar{q}_3$ is the synthetic data sample moment averaged over three implicates, and Q is the population moment. For quantiles, relative bias is $(\bar{q}_3 - Q)/\mu$ where μ is the population mean. There is no detectable bias in any of these quantities for synthetic data imputed using the correct (exact) model. Our density-based transformation also performs well: on average, there is slight reduction in skewness and kurtosis of y_1 and y_2, but no consistent bias in other quantities. In contrast, synthetic data based on additive spline regression exhibit significant bias in all quantities except the mean and median.

To investigate the repeated sampling properties of the synthetic data, we calculate 95 percent confidence interval coverage for a large number of estimands in the observed and synthetic data. Estimands include the means of all three confidential variables by group; all bivariate product-moment and rank-order correlations by group; and coefficients in the linear regression of $\ln y_2$ on an intercept, x_1, x_2, and $\ln y_1$ by group. All inferences are based on the methods of Section 2.1; confidence intervals for correlations are based on Fisher's z transformation. Figure 2 plots confidence interval coverage in the observed data versus coverage in the synthetic data. In all but a few cases, synthetic data generated using the exact model and our density-based transformation yield inferences very similar to the observed data. Synthetic data based on spline regression exhibit large distortions for some estimands.

We investigate both attribute and identity disclosure risk in the synthetic data. Our assessment of attribute disclosure risk follows Reiter (2005d). We assume an intruder estimates unit i's value of the k^{th} confidential variable, $y_{k,i}$, by averaging the unit's synthetic values across all three implicates: $\overline{\tilde{y}}_{k,i} = \sum_{m=1}^{3} \tilde{y}_{k,i}^m$ for $k = 1, 2, 3$. This is conservative, in the sense that it assumes that the intruder can identify which record in each synthetic implicate corresponds to the same source record in the confidential data. Intruders are unlikely to have this information in practice. We calculate the relative root mean squared error ($RRMSE$)

Table 1: Attribute disclosure risk (RRMSE) in Simulated Data

Method	Variable	Min	1^{st} Percentile	1^{st} Quartile	Median
Exact	y_1	.01	.07	.25	.38
	y_2	.01	.06	.21	.31
Density Transform	y_1	.01	.07	.26	.39
	y_2	.01	.06	.21	.32
	y_3	.01	.05	.25	.49
Spline Regression	y_1	.01	.05	.26	.43
	y_2	.01	.06	.24	.49
	y_3	.01	.04	.20	.39

of this estimator for each unit:

$$RRMSE_{k,i} = \left(\sqrt{\left(y_{k,i} - \overline{\tilde{y}}_{k,i}\right)^2 + M^{-1}\left(M-1\right)^{-1} \sum_{m=1}^{M} \left(\tilde{y}^m_{k,i} - \overline{\tilde{y}}_{k,i}\right)^2} \right) / y_{k,i}.$$

The distribution of $RRMSE$ in the synthetic data provides a measure of variability in the imputations. Table 1 presents averages over 5000 simulations of quantiles of the distribution of $RRMSE$. All three synthesis methods perform similarly. On average, the minimum relative root mean squared error of prediction is about 1% for all three confidential variables under all three methods. Median $RRMSE$ varies between 31% and 49%, which suggests there is a wide range of uncertainty in the imputations for most units.

We assess the risk of identity disclosure via re-identification, as suggested by Elliot (2001), Domingo-Ferrer and Torra (2003), Winkler (2004) and others. Re-identification simulates the behavior of an intruder who attempts to determine respondent identity by matching released data to a secondary data source on the basis of common variables. A re-identification simulation uses record-linkage techniques to match records in the partially synthetic data to the underlying confidential data. A partially synthetic record is deemed at high risk of identity disclosure if it is matched to its confidential source record. Re-identification provides a conservative assessment of identity disclosure risk, because it assumes that the intruder has the maximum possible information available to identify records in the synthetic data: the confidential data themselves.

Our re-identification experiment takes the arguably conservative approach of averaging synthetic values of each confidential variable across the three implicates in each replication.[11] Then, in each of the 50 cells of the cross-classification of the disclosable variables ($g \times x_1 \times x_2$), we calculate the Mahalanobis distance between each averaged synthetic record and each observed data record. The closest observed data record to a synthetic record constitutes a match.[12] If a synthetic record is matched to its source record, the record is deemed re-identified.

The overall re-identification rate is very low. Averaged over 5000 simulations, the overall re-identification rate is 0.5 percent when using the exact synthesis model or our density-based method, and 0.8 percent in synthetic data based on spline regression. Across cells, the re-identification rate corresponds closely to the inverse of cell size (see Figure 3). The inverse of cell size is a natural lower bound on the re-identification rate, since the expected number of re-identifications per cell is one if synthetic records are matched to confidential records at random.[13] Figure 3 shows that re-identification rates in synthetic data based on the exact model and our density-based method are very close to this lower bound, and somewhat larger in synthetic data based on splines.

Our density-based synthesis method and synthesis based on additive splines had similar computational demands in the reported simulations: both averaged about 23 seconds to synthesize all three variables in each replication, versus 6.5 seconds for synthesis using the exact model. Additional simulations indicate that execution time scales linearly with the number of observations for all three methods. However, synthesis based on splines is much

[11] As noted previously, averaging across implicates assumes that the intruder can identify which records in the MIPS data correspond to the same source record in the confidential data. This is a conservative assumption, because intruders are unlikely to have this information in practice. Conditional on being able to combine record-level information across implicates, however, we note that simple averaging is not necessarily optimal behavior on the intruder's part.

[12] We are implicitly assuming that the intruder knows the disclosable variables are not synthetic and therefore requires exact agreement on the disclosable variables.

[13] Domingo-Ferrer and Torra (2003) show that if two files contain n records on the same set of n respondents, the probability of correctly re-identifying exactly r respondents using a random matching strategy is $p(r) = \frac{1}{r!}\sum_{v=0}^{n-r}(-1)^v/v!$. It follows that the expected value of r is 1 for any n, or equivalently, the probability that a randomly selected record is re-identified is $1/n$.

less efficient than the other methods as the number of covariates increases. In further simulations, introducing three additional covariates in the synthesis models had no noticeable effect on execution time for synthesis via the exact method or our density-based method, but increased execution time by a factor of ten for synthesis based on splines.

5 Application

We apply our density-based transformation, coupled with linear regression, to synthesize earnings and date of birth in the Longitudinal Employer-Household Dynamics (LEHD) Program database. These are confidential administrative data based on the universe of quarterly employment records collected by state agencies to administer the Unemployment Insurance (UI) system. The LEHD database integrates the UI employment reports with a variety of internal Census Bureau data sources to attach individual and employer characteristics to the administrative records. See Abowd et al. (2004) for a detailed description of the LEHD data. We select a simple random sample of individuals employed in one state (whose identity is confidential) between 1990 and 1998.[14] The sample contains about 30 million quarterly employment records on about 1.25 million individuals.

5.1 Synthesis Details

We produce three synthetic implicates of reported earnings and date of birth. We synthesize these variables sequentially, with earnings following date of birth. For each variable, the synthesis procedure follows Example 1.

Date of birth is integer-valued and reported with daily detail. Earnings are reported quarterly in dollars. We treat both distributions as continuous. To synthesize date of birth, $\mathbf{W}_1$ includes sex, race, county of residence, and several indicators for missing data. To synthesize earnings, $\mathbf{W}_1$ includes sex, race, industry of employment (SIC Major Division),

[14] We cannot disclose the sampling rate for confidentiality reasons.

an indicator for quarterly earnings over $100,000, indicators for full-time employment and foreign birth, and several indicators for missing data.[15]

We apply our density-based transformation in each cell of the cross-classification of variables in $\mathbf{W}_1$ that contains sufficient data (at least ten times as many observations as predictors in the regression model). We collapse small cells, in which case we add main effects for the collapsed cells to $\mathbf{W}_2$.[16] Transformations are based on integrated kernel densities, estimated on an approximate Bayesian bootstrap sample of observations in each cell. In our imputation model for earnings, we apply a similar transformation to up to two leads and lags of earnings at the same employer (where these exist).

To synthesize date of birth, $\mathbf{W}_2$ includes a quartic in years of education, an indicator for foreign birth, the number of quarters worked in each year, the proportion of employment spells that were full-time, the proportion of employment spells in each SIC Major Division and county, and the mean and variance of (log) firm size and payroll in the individual's employment history. To synthesize earnings, $\mathbf{W}_2$ includes a quartic in age (based on reported/synthetic date of birth), up to two transformed leads and lags of earnings at the same employer (where these exist), a quartic in years of education, main effects for county of residence and county of employment, main effects for non-employment in each year of the sample, the employer's (log) employment and payroll, and main effects for year and quarter.

Many predictors in $\mathbf{W}_2$ are highly correlated, so we apply a simple model selection rule to increase precision in the posterior distribution of regression coefficients. On each subdomain, we estimate a candidate regression on all elements of $\mathbf{W}_2$. Only those variables that meet the Schwarz (1978) criterion are retained. We then estimate the final imputation model on the reduced set of predictors.

[15] Date of birth and earnings imputations condition on different variables because earnings varies over time, but date of birth does not. Using different predictors to generate the synthetic values implies conditional independencies in the synthetic data that depend on imputation order.

[16] The cross-classification of variables in $\mathbf{W}_1$ defines over 100,000 cells for each variable. Most of these are sparsely populated, which necessitates collapsing many small cells. Although only about ten percent of observations are in collapsed cells, cell collapse reduces the number of cells below 1,000 for date of birth, and below 3,000 for earnings. Cell sizes vary between approximately 1,500 and 1.4 million observations. The median cell size is approximately 3,150 observations.

Table 2: Moments and Quantiles of Marginal Distributions

	Age on Jan 1, 1990		Quarterly Earnings	
	Observed	Synthetic	Observed	Synthetic
Mean	29.7	29.7	6,779	6,685
Standard Deviation	16.3	16.4	22,930	26,223
Skewness	.53	.49	433	490
Kurtosis	−.10	−.11	632,771	640,980
1^{st} Percentile	4.0	3.4	43	42
5^{th} Percentile	7.4	7.2	229	242
Median	28.2	28.4	4,653	4,593
95^{th} Percentile	59.9	59.8	18,341	18,781
99^{th} Percentile	72.3	72.2	38,481	41,941
N	1,288,324		29,991,540	

5.2 Results

Table 2 reports moments and quantiles of the marginal distributions of age and earnings in the observed and synthetic data. The synthetic data replicate all of these quantities closely. The synthetic data are slightly more dispersed than the observed data. The distribution of synthetic age is slightly more symmetric and has slightly thinner tails than observed age, and the reverse is true for employment earnings.

The distributions of age and earnings are also well replicated on subdomains of $\mathbf{W}_1$. Figure 4 plots the estimated densities of observed and synthetic age by race, and Figure 5 does the same for earnings in the range \$1 to \$40,000 (which exceeds the 99^{th} percentile of the distribution of observed data).[17] In each case, the synthetic data accurately reproduce the shape of the distribution of observed data, including the number and location of modes, tail thickness, etc., although the synthetic densities are somewhat smoother. Smoothing arises because the synthetic densities are averaged over three implicates in the plots, because we collapse some small cells, and because our transformations are based on kernel density estimates. If desired, this could be mitigated by choosing a smaller bandwidth for kernel density estimates.[18]

[17] Plots by sex and race are appendicized. Plots on other subdomains are available on request.

[18] Throughout, bandwidth selection is based on Silverman's (1986) rule of thumb.

Table 3: Moments of Observed and Synthetic Quarterly Earnings

	Below $100,000		Above $100,000	
	Observed	Synthetic	Observed	Synthetic
Mean	6,343	6,420	245,934	251,665
Standard Deviation	7,414	7,851	448,977	533,795
Skewness	3.9	4.1	30.0	29.6
Kurtosis	27.3	29.3	2,298	1,849
N	29,936,981	29,937,054	54,559	54,486

The observed distribution of earnings is very right-skewed. This feature of the distribution is replicated in the synthetic data – in part because we included an indicator for quarterly earnings above $100,000 in $\mathbf{W}_1$. Table 3 reports moments of observed and synthetic earnings above and below this value. There is some small upward bias in the mean and variance on both subdomains. The synthetic distribution is slightly more skewed and has thicker tails below $100,000; the reverse is true above $100,000. Overall, however, synthetic and observed data moments are very similar.

To assess whether the synthetic data yield valid inferences about quantities of interest, we investigate their repeated sampling properties. Treating the LEHD data as a population, we take 1,500 simple random samples of 50,000 observations and calculate a variety of estimands on the observed and synthetic data. Inferences are based on the methods of Section 2.1. We use the finite population correction factor in determining the variance of all estimands.

Table 4 summarizes the repeated sampling properties of various sample means, proportions, and correlations. For most estimands, the average of synthetic data point estimates in repeated samples is close to the corresponding population value. The median ratio of the mean squared error of the synthetic data point estimate over the mean squared error of the observed data estimate is 1.05, which indicates that most synthetic and observed data point estimates are similar. Observed and synthetic confidence interval coverage are also close for most estimands, indicating that the synthetic and observed data yield similar inferences about population quantities. The only notable discrepancy is that some correlations are slightly attenuated in the synthetic data.

Table 4: Repeated Sampling Properties of Simple Estimands

Estimand	Population Value	Avg. Synthetic Estimate	95% CI Coverage	
			Observed	Synthetic
Mean age on Jan 1, 1990	31.5	31.8	95.2	93.7
Proportion age > 60 on Jan 1, 1990	.049	.044	94.0	92.5
Proportion age < 25 on Jan 1, 1990	.355	.360	95.1	98.2
Mean quarterly earnings (QE)	6,779	6,865	91.0	97.2
Proportion QE $>$ \$50,000	.006	.007	91.5	98.3
Proportion QE $<$ \$3,000	.381	.387	94.6	95.8
Mean annual earnings, all jobs (AE)	29,374	29,991	92.6	94.6
Proportion AE $>$ \$100,000	.026	.027	94.0	96.4
Correlation between age and:				
Quarterly earnings	.117	.110	95.0	83.6
Years of education	.154	.142	93.9	74.8
Indicator for foreign birth	.039	.044	96.3	93.0
ln(employer size)	.059	.059	95.1	99.6
ln(employer payroll)	.095	.091	95.3	95.3
Correlation between QE and:				
Years of education	.117	.112	95.3	86.1
Indicator for foreign birth	$-.003$	$-.002$	94.3	98.9
ln(employer size)	.053	.051	95.4	94.6
ln(employer payroll)	.117	.112	95.3	85.3
One quarter lagged QE	.410	.403	94.8	81.2
Two quarters lagged QE	.387	.390	94.8	81.4

Note: Population mean age differs from Table 2 because we sample employment records (not individuals) with equal probability.

We calculate the sample means and proportions reported in Table 4 on subdomains of sex, race, and sex × race; and, only for estimands based on earnings, on subdomains of SIC Major Division (we do not do likewise for age because industry is a characteristic of a job, not an individual). Point estimates based on the synthetic and observed data are similar on these subdomains. Over estimands and subdomains, the median ratio of the MSE of the synthetic data point estimate over MSE of the observed data estimate is 1.16. Figure 6 plots 95 percent confidence interval coverage in synthetic data versus observed data for these estimands. Coverage is similar in most cases, though somewhat greater in the synthetic data: median coverage is 94.0 in the observed data and 96.2 in the partially synthetic data. Observed data intervals frequently under-cover population quantities, particularly for earnings estimands, and this is not always reproduced in the synthetic data. Likewise, some synthetic data intervals (particularly the proportion over age 60) under-cover their population counterpart on subdomains, despite good coverage in the observed data. These discrepancies likely reflect the relatively small number of synthetic implicates. Overall, however, the observed and synthetic data yield very similar inferences about most quantities on these subdomains.

Table 5 presents estimates of a regression model of substantive interest. The model predicts the natural logarithm of quarterly earnings based on individual and employer characteristics for a sample of men employed full time.[19] This is a very well-studied specification. On the whole, the observed and synthetic data yield very similar inferences. The estimated experience profile, which is of interest to labor economists, is virtually identical in the two databases.[20] The only notable discrepancies are in several of the industry main effects. Confidence intervals in the observed data significantly undercover the population coefficients on

[19]The estimated specification differs from our synthesis model for earnings. It is based on log earnings, instead of the density-based transformation. It includes a quartic in labor force experience (which is a function of age), rather than age. Furthermore, the estimated specification includes main effects for foreign birth and the employer's industry, whereas these variables were in $\mathbf{W}_1$ for synthesis; and excludes leads and lags of earnings, main effects for county of residence and employment, and main effects for non-employment in each year.

[20]Labor force experience is a function of age. In the first period that an individual appears in the LEHD data, experience equals age minus years of education minus six. Experience increments by .25 in each subsequent quarter of employment.

employer size and payroll, and this is accurately reflected in the synthetic data. Indeed, confidence interval coverage is very similar in the observed and synthetic data for most coefficients.

We do not attempt to assess identity disclosure risk in the partially synthetic data, because synthesizing only these two variables is almost certainly insufficient to prevent re-identification.[21] We therefore focus on attribute disclosure risk. As in Section 4, we assume an intruder attempts to predict the value of a confidential variable by averaging synthetic values across implicates. Table 6 summarizes the distribution of the $RRMSE$ of this estimator in the synthetic data. For confidentiality reasons, we can not report minimum values. Instead, we report the proportion of cases with $RRMSE \leq 0.02$, and selected quantiles of the distribution of $RRMSE$. Overall, there is considerable uncertainty about observed values of age and earnings. The greatest risk of attribute disclosure arises from aggregating individual earnings across all jobs in a calendar year (AE). Even here, less than 0.5 percent of cases have $RRMSE$ below 2 percent, and $RRMSE$ exceeds 17 percent for the median observation, so there is significant uncertainty about true earnings. Median values of $RRMSE$ are more than twice this large for age and quarterly earnings.

Data collectors are likely to be particularly concerned about attribute disclosure risk when individuals have extreme values of age or earnings. The lower panel of Table 6 summarizes the distribution of $RRMSE$ in these potentially sensitive cases. Uncertainty is even greater for those with extreme values of earnings than in the population as a whole, and remains large for extremes of age. This suggests the partially synthetic data provide strong protection against attribute disclosure, even for extreme cases.

6 Conclusion

Statistical disclosure limitation methods promise high quality microdata with low disclosure risk. Among existing disclosure limitation methods, multiply-imputed partially synthetic

[21] The large number of unsynthesized variables on the file will be sufficient to re-identify many records.

Table 5: Estimated Coefficients in Log Earnings Regression

	Population Value	Avg. Synthetic Estimate	95% CI Coverage	
			Observed	Synthetic
Years of experience	.086	.086	91.7	95.1
Experience^2/100	−.321	−.327	89.2	94.4
Experience^3/1000	.055	.057	86.9	94.9
Experience^4/10000	−.004	−.004	84.8	95.2
Initial Experience < 0	−.176	−.167	94.6	94.1
Years of Education	.012	−.002	94.1	97.8
Education^2/100	−.541	−.461	93.9	98.3
Education^3/1000	.733	.783	94.3	98.2
Education^4/10000	−.196	−.226	94.7	97.7
Race = Black	−.271	−.264	96.5	98.7
Race = Hispanic	−.205	−.186	97.3	92.3
Foreign born = 1	−.078	−.054	95.1	91.1
ln(Employer size)	−.372	−.414	48.4	47.3
ln(Employer payroll)	.397	.440	43.7	42.8
SIC Division = A	−.136	−.164	92.3	97.1
SIC Division = B	−.059	−.058	97.6	99.6
SIC Division = C	.031	.008	95.5	97.1
SIC Division = E	−.005	−.010	98.2	99.7
SIC Division = F	.041	.020	98.3	97.4
SIC Division = G	−.208	−.192	88.3	91.5
SIC Division = H	.089	.061	95.7	88.3
SIC Division = I	−.211	−.244	92.7	75.1
SIC Division = J	−.218	−.237	94.3	95.3
Year = 1991	−.003	.000	95.7	97.8
Year = 1992	.025	.023	97.3	99.3
Year = 1993	.040	.044	96.4	98.5
Year = 1994	.068	.069	95.2	98.3
Year = 1995	.081	.080	96.7	98.8
Year = 1996	.104	.102	95.9	98.8
Year = 1997	.126	.124	95.5	98.5
Year = 1998	.152	.148	95.1	98.4
Quarter = 2	.053	.051	93.9	97.1
Quarter = 3	.047	.048	93.5	97.9
Quarter = 4	.079	.089	94.3	96.4
Intercept	4.20	4.00	71.7	74.6
RMSE	.762	.864		
Number of Observations	7,145,344	11,910		

Table 6: Attribute Disclosure Risk (RRMSE) in Partially Synthetic Data

Variable	Proportion ≤ 0.02	1^{st} Percentile	1^{st} Quartile	Median
Age on Jan 1, 1990	.002	.079	.286	.441
Quarterly earnings (QE)	.001	.049	.198	.361
Annual earnings, all jobs (AE)	.005	.026	.101	.174
Potentially Sensitive Cases:				
Age on Jan 1, 1990 ≥ 60	.002	.031	.227	.385
QE $\leq \$1000$	.001	.192	.714	1.89
QE $\geq \$100,000$	.001	.058	.252	.423
AE $\geq \$500,000$	.001	.065	.292	.501

data strike a compelling balance between these competing objectives. Indeed, the main virtue of this approach is that it preserves the ability of users to obtain valid statistical inferences about a population of interest. Our simulation and application to LEHD data demonstrate the high utility and low disclosure risk of MIPS data. In simulations, our method of generating synthetic values delivered better data utility and lower disclosure risk than a nonparametric alternative, at lower computational cost. Our application to LEHD data demonstrates that our approach is feasible in large scale applications and performs well in genuine data.

Like all model-based disclosure limitation methods, the quality of MIPS data depends on correctly specifying the imputation model. Our transformation-based methods mitigate mis-specification that can arise when the correct imputation model is unknown. Mis-specification is still possible, because MIPS data will only preserve multivariate relationships reflected in the imputation model. To preserve *all* multivariate relationships in the partially synthetic data requires, in principle, that the imputation model conditions on "everything." This is not possible in practice. We saw evidence of this in our application to LEHD data, where it was necessary to collapse some subdomains on which we sought to preserve the conditional distribution of age and earnings, and to reduce the number of predictors in imputation regressions though model selection. Further research is required to determine optimal methods for reducing the dimensionality of the synthesis problem.

By definition, all model-based imputation methods impose some degree of modeling burden. Our method partly alleviates this burden, because data collectors need only define subdomains of primary and secondary interest and specify a simple imputation model that captures relationships of secondary interest within subdomains. This is less onerous than completely specifying a parametric imputation model for $\mathbf{y}_k|\mathbf{W}$. In cases where data collectors are unable to specify a credible parametric model for $\mathbf{z}_k|\mathbf{W}_2$, our density-based transformations could be combined with nonparametric regressions on subdomains of primary interest. We leave this for future research.

It is important that data collectors recognize and advertise the limitations of partially synthetic data they release. In particular, the model used to generate the MIPS data will make them well suited to some analyses and poorly suited to others. Data collectors must therefore accompany any release of MIPS data with sufficient information for users to determine whether the released data are appropriate for their analysis.

References

Abowd, J. M., J. Haltiwanger, and J. Lane (2004). Integrated longitudinal employer-employee data for the United States. *American Economic Review 94*(2), 224–229.

Abowd, J. M. and S. D. Woodcock (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. I. Lane, J. J. Theeuwes, and L. V. Zayatz (Eds.), *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, Chapter 10, pp. 215–278. North-Holland.

Domingo-Ferrer, J. and V. Torra (2003). Disclosure risk assessment in statistical microdata protection via advanced record linkage. *Statistics and Computing 13*, 343–354.

Duncan, G. T., S. A. Keller-McNulty, and S. L. Stokes (2001). Disclosure risk vs. data utility: The r-u confidentiality map. National Institute of Statistical Sciences Technical Report No. 121.

Duncan, G. T. and D. Lambert (1986). Disclosure-limited data dissemination. *J. American Statistical Association 81*(393), 10–18.

Elliot, M. (2001). Disclosure risk assessment. In P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. M. Zayatz (Eds.), *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, Chapter 4, pp. 75–90. North-Holland.

Fienberg, S. E. (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Carnegie Mellon University Department of Statistics Technical Report No. 611.

Fienberg, S. E. and U. E. Makov (1998). Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics 14*(4), 385–397.

Fienberg, S. E., U. E. Makov, and R. J. Steele (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics 14*(4), 485–502.

Kennickell, A. B. (1997, November). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. SCF Working Paper.

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics 9*(2), 313–331.

Lin, L. I.-K. and E. F. Vonesh (1989). An empirical nonlinear data-fitting approach for transforming data to normality. *The American Statistician 43*(4), 237–243.

Little, R. and F. Liu (2003). Selective multiple imputation of keys for statistical disclosure control in microdata. The University of Michigan Department of Biostatistics Working Paper Series.

Muralidhar, K. and R. Sarathy (2003). A theoretical basis for perturbation methods. *Statistics and Computing 13*, 329–335.

Nusser, S., A. Carriquiry, K. Dodd, and W. Fuller (1996). A semiparametric transformation approach to estimating usual daily intake distributions. *Journal of the American Statistical Association 91*(436), 1440–1449.

Raghunathan, T., J. Reiter, and D. Rubin (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics 19*(1), 1–16.

Raghunathan, T. E., J. M. Lepkowski, J. V. Hoewyk, and P. Solenberger (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology 27*(1), 85–95.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics 18*(4), 531–544.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology 29*, 181–188.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology 30*, 235 – 242.

Reiter, J. P. (2005a). Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association 100*(472), 1103–1112.

Reiter, J. P. (2005b). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*, 185–205.

Reiter, J. P. (2005c). Significance test for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference 131*, 365–377.

Reiter, J. P. (2005d). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics 21*, 441–465.

Reiter, J. P. and T. E. Raghunathan (2007, December). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association 102*, 1462–1471.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. B. (1993). Discussion of statistical disclosure limitation. *Journal of Official Statistics 9*(2), 461–468.

Sarathy, R., K. Muralidhar, and R. Parsa (2002). Perturbing nonnormal confidential attributes: The copula approach. *Management Science 48*(12), 1613–1627.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

Willenborg, L. and T. de Waal (1996). *Statistical Disclosure Control in Practice*. Springer-Verlag.

Winkler, W. E. (2004). Re-identification methods for masked microdata. In J. Doming-Ferrer and V. Torra (Eds.), *Privacy in Statistical Databases*, pp. 216–230. Springer. Lecture Notes in Computer Science 3050.

Figure 1: Simulated Bias in Moments and Quantiles

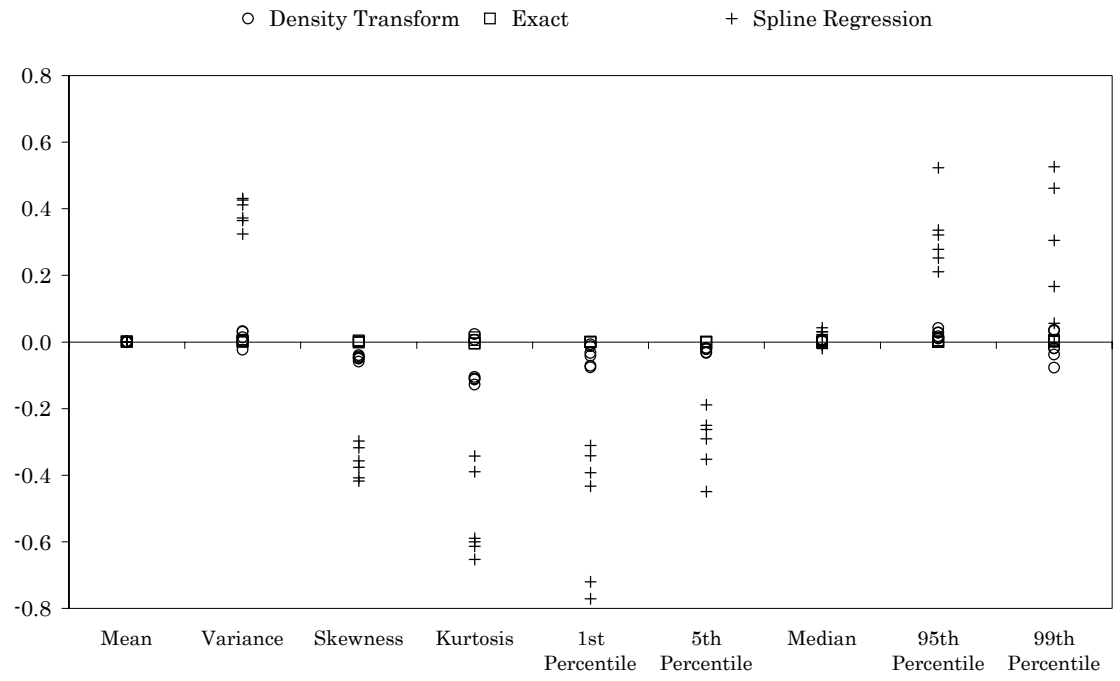
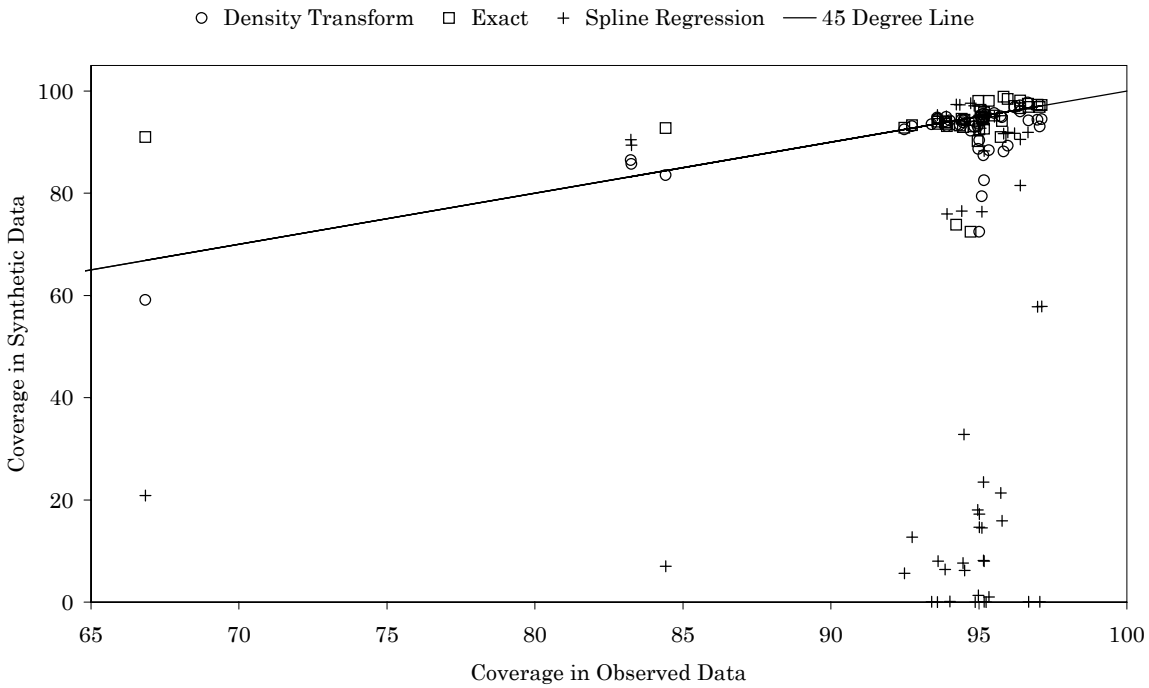


Figure 2: Simulated 95% Confidence Interval Coverage



Note: Estimands include means of y_1, y_2, y_3 by g; product–moment and rank–order correlations between (y_1,y_2,y_3,x_1,x_2) by g; and coefficients in the linear regression of $\ln y_2$ on an intercept, x_1, x_2, and $\ln y_1$ by g.

Figure 3: Simulated Re-identification Rates by Cell

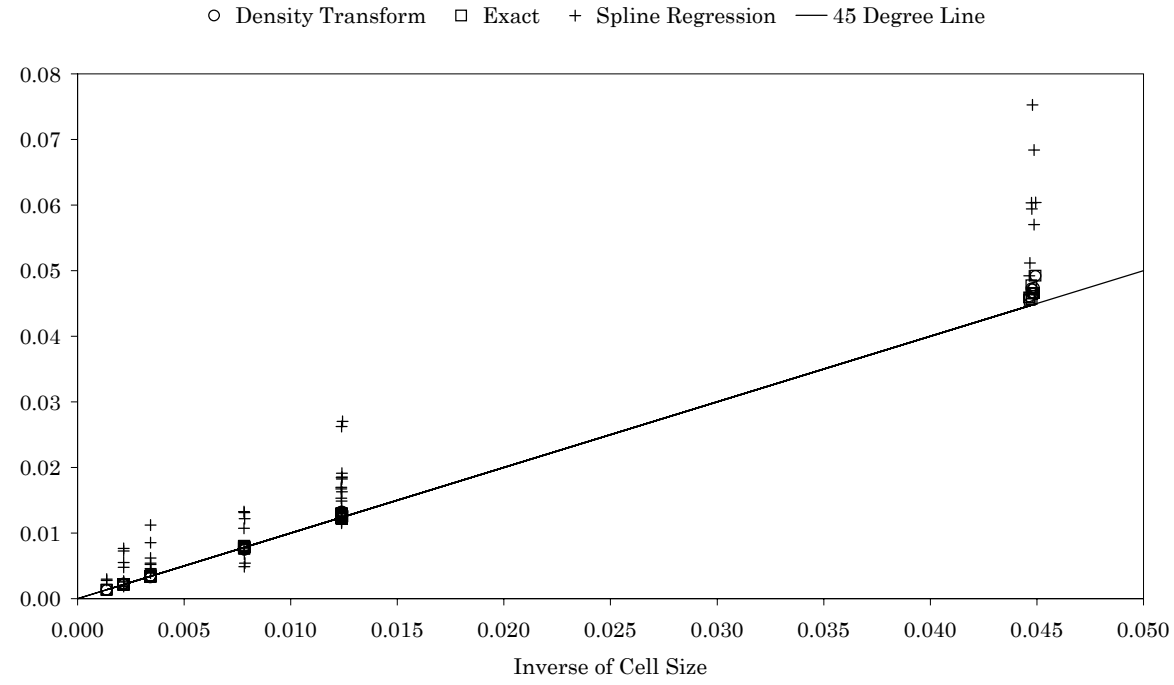


Figure 6: 95% Confidence Interval Coverage of Sample Means and Proportions on Sex, Race, and Industry Subdomains

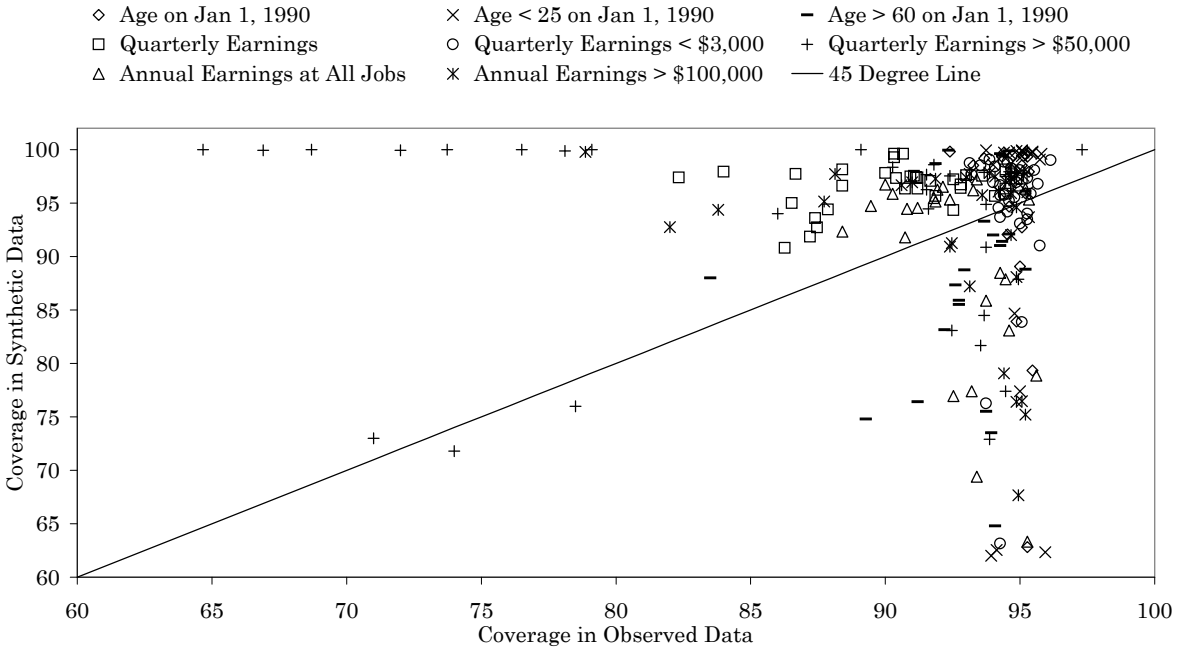


Figure 4
Estimated Density of Observed and Synthetic Age on Jan 1, 1990

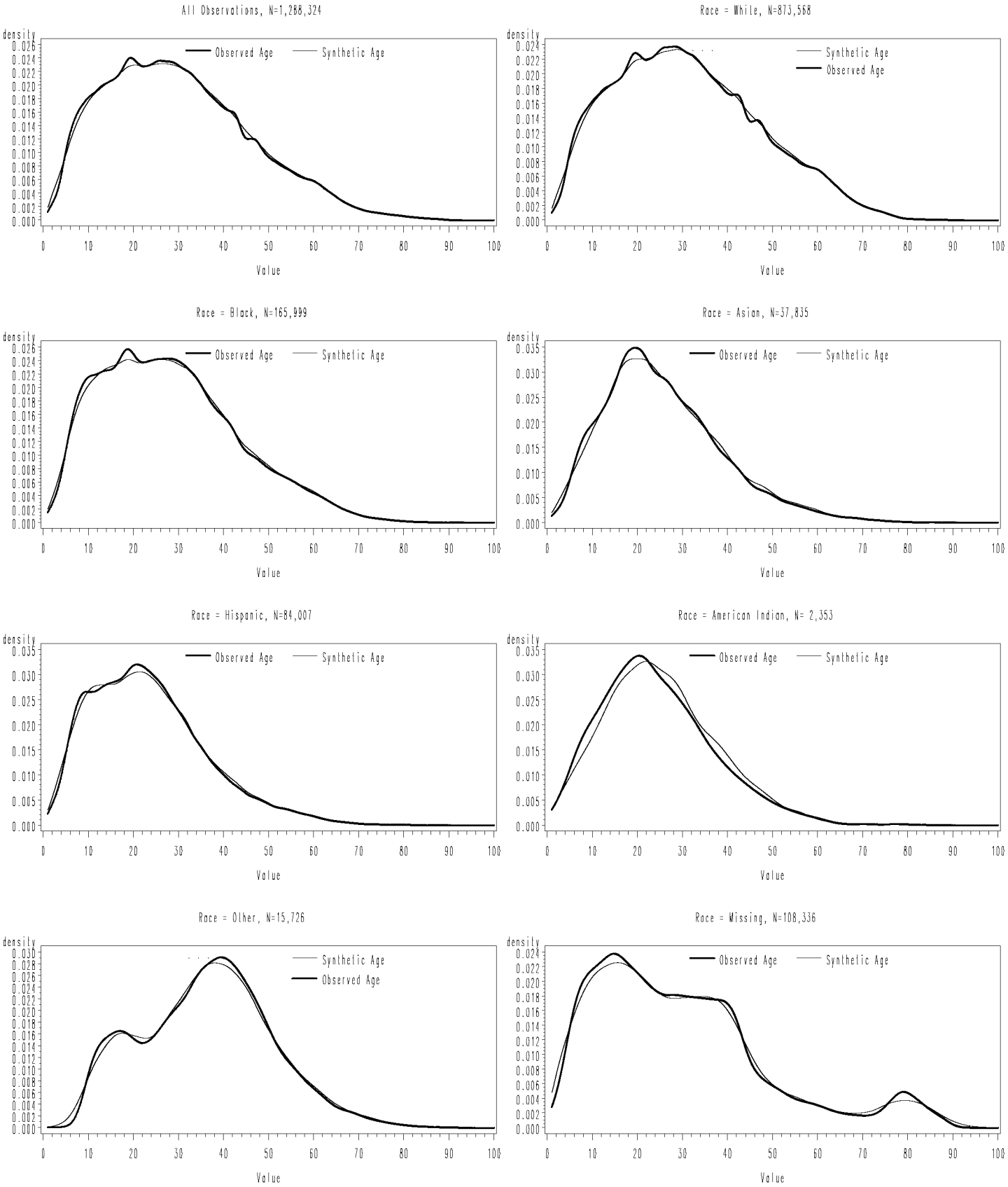


Figure 5
Estimated Density of Observed and Synthetic Quarterly Earnings